Mathematics Education Research Journal 2006, Vol. 18, No. 2, 40-55

A Longitudinal Study of Student Understanding of Chance and Data

Jane Watson
University of Tasmania

Ben Kelly
University of Tasmania

John Izard
RMIT University

This study uses Partial Credit Rasch analysis to study a complex data set of student responses to survey items relating to chance and data. The items were administered in the classroom and collected from 1993 to 2003 in the Australian state of Tasmania. Data were collected from a total of 5514 individual students across Grades 3 to 11 over the decade and of these students 896 provided at least one repeated measure. As students completed a core of common items, Rasch analysis could be performed and all students were subsequently placed on the same logit scale for comparison. The purpose of the analysis is to consider average cohort change over time and trends in performance during the first 10 years after the curriculum was introduced in Tasmania. Implications for the education system and curriculum implementation are considered.

The topics Chance and Data were officially introduced into the mathematics curriculum in the Australian state of Tasmania in 1993 in the *Mathematics Guidelines K-8* (Department of Education and the Arts, 1993). This followed their introduction in the United States by the National Council of Teachers of Mathematics (NCTM, 1989) and in Australia in *A National Statement on Mathematics for Australian Schools* (Australian Education Council [AEC], 1991). From 1993 a series of projects followed cohorts of Tasmanian school students over the next decade, identifying understanding, and using surveys that were classroom-administered tests, which students were told were not associated with their school assessment but for research only. The purposes of the research included constructing models for the development of understanding of chance and data concepts. The survey data collected throughout the projects also allowed the comparison of cohorts across the first decade following the implementation of the Chance and Data curriculum in Tasmania. The comparison gave an indication of the success of the curriculum in increasing student understanding of the topics.

Background

Before the introduction of the NCTM's *Standards* in 1989 there was little research into school students' understanding of statistical concepts. Green (1983, 1986), Fischbein (1975), and Piaget and Inhelder (1951/1975) were the major contributors in the area of probability and Goodchild (1988), Mevarech (1983), and Strauss and Bichler (1988) on the topic of average. The structure of the new curriculum (following, e.g., Holmes, 1980), however, was more broadly based, focussing on all aspects of a statistical investigation: data

collection and sampling, data representation, data reduction, probability, and analysis and inference. The initial research in Tasmania based on a representative sample of schools hence set out to cover all components of the curriculum (Watson, 1994). Reports on each of the five aspects identified in the new curriculum have appeared in recent years from the Tasmanian studies and others around the world (e.g., Cai, 1998; Friel, Curcio, & Bright, 2001; Jacobs, 1999; Lehrer & Romberg, 1996; Shaughnessy, 2003; Watson, 2006).

As well as an emphasis on the statistical investigation, from 1990 statisticians and statistics educators stressed the importance of variation as the phenomenon underlying statistical investigations (Moore, 1990; Wild & Pfannkuch, 1999) and the need for educational research in this area (Green, 1993; Shaughnessy, 1997). This prompted a project in Tasmania in a group of schools that had not been involved in earlier studies that not only explored student understanding of variation (e.g., Kelly & Watson, 2002; Watson & Kelly, 2004a, 2005) but also provided instruction in Chance and Data that emphasised variation (e.g., Watson & Kelly, 2002). The surveys in the variation project included items from earlier surveys as well as items dealing explicitly with variation as it occurs at various stages of a statistical investigation. Watson, Kelly, and Izard (2004) reported on the survey outcomes of this variation project with overall results indicating that, although eight weeks after instruction mean scores improved, after two years there was little difference between the average outcomes for students in the schools that had experienced the special lessons and the schools that had experienced the usual curriculum with no intervention from the project.

During this decade (1993 to 2003) wider issues of statistical literacy were canvassed in the education community, both for school students and adults (Gal, 2002; Wallman, 1993; Watson, 1997). Data from 1993, 1995, 1997, and 2000 were used by Watson and Callingham (2003) to suggest a developmental pathway of understanding of statistical literacy. Using Rasch techniques they described six levels of increasing facility with statistical ideas. At the Idiosyncratic Level (Level 1), students can read cells in tables and carry out one-to-one counting tasks but are likely to produce tautologies or other responses unrelated to the tasks presented. At the Informal Level (Level 2), they carry out one-step calculations but generally express intuitive beliefs for example about probability. At the Inconsistent Level (Level 3), students are likely to give qualitative rather than quantitative responses to tasks and show a limited appreciation of content and context. At Level 4, Consistent Non-Critical, they show straightforward engagement with context and generally deal with simple means, probabilities and graphs. At Level 5, Critical, students appear to appreciate the part variation plays in most contexts and are able to question claims that do not require mathematical skills, particularly proportional reasoning. Finally at the Critical Mathematical Level (Level 6), students engage in critical questioning of tasks, employing proportional reasoning and nuanced language in responses.

In 2003 it was possible to survey students in the schools that had participated in the original project a decade earlier. A survey was devised comprised of items across the Chance and Data curriculum from the earlier

Tasmanian studies and including items addressing consideration of the underlying concept of variation from the variation study. This made it possible to compare across all years in which surveys were conducted. Initially comparisons were made only for schools in which original and longitudinal data were collected and the hierarchy of statistical literacy understanding suggested by Watson and Callingham (2003) was confirmed (Watson, Kelly, & Izard, 2005).

The use of partial credit Rasch analysis (Masters, 1982) allows the comparison of all cohorts over the decade due to the use of common items linking across the surveys. This type of analysis has not occurred previously in relation to understanding of Chance and Data and has the potential to identify concepts that are stable or unstable over time. The data collected hence allow consideration of the following research questions.

1. Considering all data collected, what is the trend in student performance over the first decade after the introduction of chance and data into the mathematics curriculum in the Australian state of Tasmania?
2. Where longitudinal data are available, how do average student understandings change over two-, four-, or six-year periods?

Details on the grade levels for which these questions are addressed are given in the Methodology.

Methodology

Sample

A total of 5514 students over the decade completed surveys that are analysed in this paper. Of these 896 completed the surveys twice and 264 completed them three times. A summary of the grades and years in which surveys were completed and the sample sizes is given in Table 1. A total of 6410 surveys was completed in 28 separate samples.

Table 1
Sample Sizes for All Student Responses by Year and Grade (Numbers in Parenthesis Indicate Students Surveyed Two or Three Times)

Grade	1993	1995	1997	2000	2002	2003	Total
3	322 (147[a])	303	237 (54[b])	176 (114[b])		189	1227
5		465 (147[a])	226	183 (102[b])	114[b]		988
6	311 (117[a])	337	233			174	1055
7			314 (147[a])	186 (135[b])	102[b]		602
8		374 (117[a])	192				566
9	392 (117[b])	371 (51[b])	105	193 (59[b])	135[b]	251 (54[b])	1447
10			297 (117[a])				297
11		118 (117[b])	51[b]		59[b]		228
Total	1025	1968	1655	738	410	614	6410

Note. [a] Students surveyed three times. [b] Students surveyed twice.

Thirteen state government schools representing all regions of the state of Tasmania participated in the surveys in 1993, 1995, 1997, and 2003. Ten new schools from the Hobart suburban area participated in the surveys in 2000 and 2002. These were chosen because they were similar to each other to provide comparison data about the effectiveness of a teaching intervention. In the current study, the data for 2000 and 2002 were combined for schools in which intervention took place and those where it did not. This was because the schools had been matched initially on socio-economic criteria but after two years there were no differences in average longitudinal performance (Watson et al., 2004). The data from Grade 11 surveys were collected from the senior secondary schools for which high schools in the project were feeder schools. As senior secondary schooling is not compulsory in Tasmania, there was a considerable drop in participation for follow-up surveys at this grade level.

Instruments

The instruments used in 1993, 1995, 1997, 2000, and 2002 are included in an appendix in Watson and Callingham (2003). Those specific to 2000 and 2002 are presented in Watson, Kelly, Callingham, and Shaughnessy (2003). The selection of items for the 2003 survey was made from the earlier ones, with the addition of five new items developed for a parallel survey used in a different school system (Callingham & Watson, 2005; Watson & Callingham, 2005). The items employed over the years 1993-2003 covered the components of the Chance and Data curriculum as found in curriculum documents (AEC, 1991; NCTM, 1989, 2000) as well as specific aspects of variation. In all years where surveys took place, students in lower grades answered fewer questions than those in higher grades. Although some items were coded on a right-wrong (1-0) basis, most were coded in a hierarchical fashion based on structure and appropriateness, with codes ranging from 0-2 to 0-5. Details of coding are found in the sources noted in this paragraph and were based on the SOLO Taxonomy of Biggs and Collis (1982) or the statistical literacy hierarchy of Watson (1997).

Analysis

The data were analysed using Rasch (1960/1980) measurement techniques. The initial analyses of data collected in 2000 established anchor values for the items in common across years 1993 to 2003, so tests including these items could be calibrated on a common scale or continuum of achievement. The year 2000 was chosen because that survey contained both general items and variation items and hence was an appropriate data set for producing anchor values.

Using anchor values from the year 2000 data, the year 2003 data were analysed and a second, more comprehensive anchor file constructed. These anchored item values were then used in a series of analyses in which common items from each study year, 1993, 1995 and 1997, were added into the data pool, to create an anchor file consisting of 31 items that met the criteria for fit.

Items that misfitted at any stage were dropped from the anchor file to ensure that the final anchor file was robust. The final statistics associated with the output of this Rasch analysis were acceptable and are reported in Appendix A. This file was subsequently used to estimate all other item difficulties and to obtain ability estimates for all students in each year.

The scaled achievement scores for each student on each test could then be used to determine the initial differences among the grades and the changes over the time between testings. The effect sizes for these differences were determined using Cohen's (1969) methodology and reported with descriptors devised by Cohen (1969) and Izard (2004). The scaled scores of all students were then used in subsequent analyses. For each grade for which data were collected during the decade, the trend in mean scores is presented. Appendix B contains the means and standard deviations for each of the 28 samples detailed in Table 1 as well as the comparisons across the years for each grade except Grade 10, which was sampled only once. The means are displayed graphically in Figure 1 to indicate the trend over grades and over the decade in relation to Research Question 1. To reinforce the suggestion of trends in growth in understanding over time for individual cohorts, the 11 samples (with numbers in parenthesis in Table 1) for which longitudinal data were collected on individual students are considered in relation to Research Question 2. Appendix C contains the means, standard deviations, and effect sizes for cohort comparisons among these 11 pairs of data sets.

Results

Research Question 1

Figure 1 displays the mean logit scores for each of the 28 samples in the study. They are displayed by increasing grade and within each grade by increasing year. An overall trend for increasing mean with grade is evident. For each grade where longitudinal data were collected, 2003 has the lowest mean score. For Grade 9 data are available for every year covered by the studies, and here 2003 had the lowest mean. The same is true for 2003 in the Grade 3 and Grade 6 data. The comparisons of pairs of years within grades are presented in Appendix B, where large differences, using Cohen's (1969) criteria, are observed for Grade 3 in a negative direction between 2003 and both 1997 and 2000, for Grade 5 a positive difference for both 2000 and 2002 in relation to 1995 and 1997, and a negative difference for Grade 9 between 2003 and 1993.

The improvement in performance from both 1993 and 1995 to 1997 and 2000 was medium in Cohen's terms for successive Grade 3 groups, as was the decline between those years and 2003. The decline in performance for successive Grade 6 groups for all three years 1993, 1995, and 1997 to 2003 was also medium, as was that for Grade 9 from 1993 to 1995, 1997, and 2000. Between 2000 and 2002 for this grade there was a medium improvement but then a drop in mean performance of this degree, again from 2002 to 2003. Except for the apparent decline in 2003, there appears to be little overall trend that would indicate steady improvement or decline across the decade.

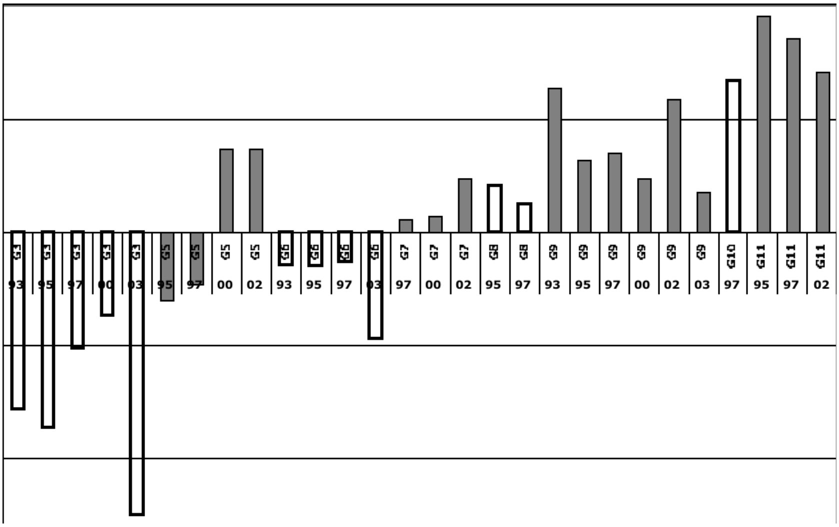

Figure 1. Mean scores for the 28 samples ordered by year within grade level.

Research Question 2

The collection of longitudinal data for some years allows the observation of cohort change over time. Given the observations for Research Question 1 it is not possible to suggest a direct curriculum implementation effect but to

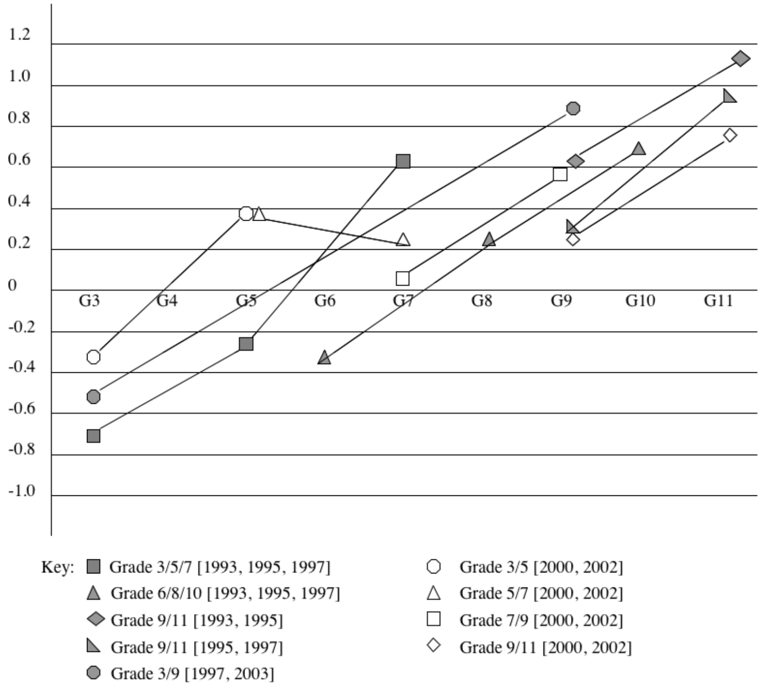

Figure 2. Relative performance for longitudinal data sets.

reinforce observations of trends of improvement with respect to maturity and general learning within school and society. Appendix C presents the means, standard deviations and comparisons for these 11 data sets and they are shown graphically in Figure 2. Lines connect pairs of data sets involving the same students. Error bars for the means have been omitted for clarity in reading the graph.

Considering the three data sets where two successive comparisons were possible for the same group of students, for students originally in Grade 3, both differences (Grade 3 to Grade 5, 1993 to 1995, and Grade 5 to Grade 7, 1995 to 1997) were large using Cohen's criteria, whereas for students originally in Grade 6, the 1993-1995 (Grade 6 to Grade 8) difference was medium and the 1995-1997 (Grade 8 to Grade 10) difference was large. In all cases there was improvement, from Grade 3 to Grade 5 to Grade 7, and from Grade 6 to Grade 8 to Grade 10. The other large difference, as might be expected, was from Grade 3 in 1997 to Grade 9 in 2003, whereas there was only a small difference for Grade 9 to Grade 11 in both 1993 to 1995, and 1995 to 1997. For the 2000 to 2002 data, the difference from Grade 3 to Grade 5 was positive and large, whereas from Grade 5 to Grade 7 it was negative and small. For Grade 7 to Grade 9 it was again positive and medium, whereas for Grade 9 to Grade 11 it was positive and small as for the earlier years.

Overall the longitudinal data suggest that the greatest improvement in understanding occurred between Grade 3 and Grade 5, whereas after a questionable period in Grades 6 and 7, again at least medium degrees of improvement occurred. Although some improvement continued to Grade 11, generally it was small compared to earlier two-year periods.

Discussion

Two aspects of this research are considered in the discussion. The first relates to the use of Rasch analysis, and the second to the observation of performance over the first decade after the introduction of the Chance and Data curriculum.

Over a decade of research, evolution in thinking occurs and researchers are influenced by other happenings in the field. So it was with this research. The original tests used with students reflected the goals of the curriculum in Chance and Data (AEC, 1991; NCTM, 1989), as well as including items used by earlier researchers (e.g., Fischbein & Gazit, 1984) and indicating the emerging interest in statistical literacy (Watson, 1997). By the end of the 1990s, however, the acknowledged importance of statistical variation meant that further items were added to tests from 2000, as well as retaining some from previous studies. The use of Rasch analysis allowed the data sets to be combined using common items in order to be able to put all students over the decade and all items on the same scale. The initial analysis of items by Watson and Callingham (2003; 2005) and Callingham and Watson (2005) indicated that the original concepts and the items focusing on variation within the original contexts produce a uni-dimensional construct of statistical understanding. The final step in this construction of instruments means that it is now possible to measure confidently students' understanding across the topics associated with statistics.

The other advantage to using Rasch analysis is related to the use of fewer survey items with younger students. In carrying out comparison studies by conventional methods using raw score totals, as was undertaken, for example, by Watson and Kelly (2004b), it is necessary to delete some items if some grades have not completed them or to do comparisons only between certain grades. It may be in such comparisons that some information is lost. Using Rasch techniques allows all students to be placed on the same scale and hence an overall picture of performance to emerge.

In terms of following the progress of students in Tasmania over the decade since the introduction of the Chance and Data curriculum the picture is mixed. Except for one group followed longitudinally from Grade 5 to Grade 7, other students' performance improved to a greater or lesser extent over the two, four, or six years they were followed. The middle school plateau has been observed by others (e.g., Callingham & McIntosh, 2002) and is a matter of concern within a context where general development appears to occur within all other groups.

In terms of growth for particular grades over the years, the primary Grades 3 and 5 show the greatest improvement over the initial years of the curriculum until 2000. Grade 9 showed mixed results over the ten years. Although it may not be realistic to expect large differences in successive years, over the decade in the original schools surveyed in 1993, the 2003 results are disappointing for Grades 3, 6, and 9.

There are some possible reasons that could be suggested including lack of continuing professional development for teachers over the decade. Certainly at the time of initial introduction there were curriculum implementation officers working in all districts in the state. These positions were discontinued in 1997 and later replaced with Literacy and Numeracy officers, whose brief was much wider than specific curriculum implementation. As well, 2000 was the beginning of the introduction of the *Essential Learnings Framework* (Department of Education, 2002) with a focus on values-based education including 18 essential elements, of which "Being Numerate" was one. Mathematics as a discipline did not feature in this framework. Emphasis on basic numeracy as part of the Essential Learnings may have reduced the focus on particular aspects of the mathematics curriculum, such as Chance and Data. No other data-driven evidence apart from this study could be found to explain the decline in performance in 2003. It would be of interest to know if a similar trend occurred in other subject areas but no such longitudinal data are known to exist.

Anecdotal evidence from a senior teacher in one of the schools involved from 1993 in the research—a school that had also been involved in a project on the theme "Thinking Mathematically" for five years—suggested that 10 years was not enough to see a change in Chance and Data. Other topics were more important and it would take 20 years for an improvement to be seen in Chance and Data, both for teachers and students.

From an analysis perspective this study illustrates the usefulness of Rasch analysis for placing students from different grades and different times

on the same scale in order to make the comparisons and for following the development of individual students. From an educational point of view the outcomes point to some concerns about achieving the goals of the curriculum implementation and about the plateau in performance for some cohorts across the middle grades.

Acknowledgments

The data used in this research were collected as part of Australian Research Council Grants No. A79532539, A00000716, and DP0208607, and University of Tasmania Grant No. W0009108. Preliminary results of some parts of the analysis were presented at conferences of the Australian Association for Research in Education (2004) and the Mathematics Education Research Group of Australasia (2005).

References

Australian Education Council. (1991). *A national statement on mathematics for Australian schools*. Melbourne: Author.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Cai, J. (1998). Exploring students' conceptual understanding of the averaging algorithm. *School Science and Mathematics, 98*, 93—98.

Callingham, R. A., & Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement, 6*(1), 19—47.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Department of Education Tasmania. (2002). *Essential learnings framework* 1. Hobart: Author.

Department of Education and the Arts. (1993). *Mathematics guidelines K-8*. Hobart: Curriculum Services Branch.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, The Netherlands: D. Reidel.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? An exploratory research study. *Educational Studies in Mathematics, 15*, 1—24.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*, 124—158.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*, 1—51.

Goodchild, S. (1988). School pupils' understanding of average. *Teaching Statistics*, 10, 77—81.

Green, D. R. (1983). A survey of probability concepts in 3000 pupils aged 11-16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (Vol. 2, pp. 766—783). Sheffield, UK: Teaching Statistics Trust.

Green, D. R. (1986). Children's understanding of randomness: Report of a survey of 1600 children aged 7-11 years. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 287—291). Victoria, BC:

The Organizing Committee, ICOTS2.

Green, D. (1993). Data analysis: What research do we need? In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: Who should teach it?* (pp. 219—239). Voorburg, The Netherlands: International Statistical Institute.

Holmes, P. (1980). *Teaching statistics 11-16.* Slough, UK: Schools Council and Foulsham Educational.

Izard, J. F. (2004, March). *Best practice in assessment for learning*. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on Redefining the Roles of Educational Assessment, South Pacific Board for Educational Assessment, Nadi, Fiji.

Jacobs, V.R. (1999). How do students think about statistical sampling before instruction? *Mathematics in the Middle School, 5(*4), 240—263.

Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 26th annual Conference of the Mathematics Education Research Group of Australasia, Auckland, NZ,Vol. 2, pp. 366—373). Sydney: MERGA.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*(1), 69—108.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*, 415—429.

Moore, D. S. (1990). Uncertainty. In L. S. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95—137). Washington, DC: National Academy Press.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Leake Jr., P. Burrell, & H. D. Fishbein, Trans.). New York: W.W. Norton. (Original work published 1951)

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960)

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *People in mathematics education* (Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia, Vol. 1, pp. 6—22). Waikato, NZ: MERGA.

Shaughnessy, J. M. (2003). Research on students' understandings of probability. In J. Kilpatrick, W.G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 216—226). Reston, VA: National Council of Teachers of Mathematics.

Strauss, S., & Bichler, E. (1988). The development of children's concept of the arithmetic average. *Journal for Research in Mathematics Education, 19*, 64-80.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, No 421, 1—8.

Watson, J. M. (1994). Instruments to assess statistical concepts in the school curriculum. In National Organizing Committee (Ed.), *Proceedings of the 4th International Conference on Teaching Statistics* (Vol. 1, pp. 73—80). Rabat, Morocco: National Institute of Statistics and Applied Economics.

Watson, J. M. (1997). Assessing statistical literacy using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107—121).

Amsterdam: IOS Press and The International Statistical Institute.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum.

Watson, J. M., & Callingham, R. A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3—46.

Watson, J. M., & Callingham, R. A. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education.* International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 2004 (pp. 116—162). Voorburg, The Netherlands: International Statistical Institute.

Watson, J. M., & Kelly, B. A. (2002). Grade 5 students' appreciation of variation. In A. Cockburn & E. Nardi (Eds), *Proceedings of the 26th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 385—392). Norwich, UK: PME.

Watson, J. M., & Kelly, B. A. (2004a). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics and Technology Education, 4*, 371—396.

Watson, J. M., & Kelly, B. A. (2004b). A two-year study of students' appreciation of variation in the chance and data curriculum. In I. Putt, R. Faragher, & M. McLean (Eds.), *Mathematics education for the third millennium: Towards 2010* (Proceedings of the 27th annual conference of the Mathematics Education Research Group of Australasia, Townsville, Vol. 2, pp. 573—580). Sydney, NSW: MERGA.

Watson, J. M., & Kelly, B. A. (2005). The winds are variable: Student intuitions about variation. *School Science and Mathematics, 105*, 252—269.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34*, 1—29.

Watson, J. M., Kelly, B. A., & Izard, J. F. (2004, December). Student change in understanding of statistical variation after instruction and after two years: An application of Rasch analysis. *Proceedings of the 2004 annual conference of the Australian Association for Research in Education*. Available at http://www.aare.edu.au/04pap/wat04867.pdf

Watson, J. M., Kelly, B. A., & Izard, J. F. (2005). Statistical literacy over a decade. In P. Clarkson, A. Downton, D. Gronn, M. Horne, A. McDonough, R. Pierce, & A. Roche (Eds.), *Building connections: Theory, research and practic*e (Proceedings of the 28th annual conference of the Mathematics Education Research Group of Australasia, Melbourne, Vol. 2., pp. 775—782). Sydney: MERGA.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223—265.

Authors

Jane M. Watson, Professor of Mathematics Education, Faculty of Education, University of Tasmania, Private Bag 66, Hobart, TAS 7001. E-mail: <Jane.Watson@utas.edu.au>

Ben A. Kelly, Faculty of Education, University of Tasmania, Private Bag 66, Hobart, TAS 7001. E-mail: <Ben.Kelly@utas.edu.au>

John F. Izard, RMIT University, GPO Box 2476V, Melbourne, VIC 3001. E-mail: <john.izard@rmit.edu.au>

Appendix A: Rasch Statistics

Summary Results: 2003 data anchored on 2000 results

Items 1to37 2003 data (Run No 8)	
Item Estimates (Thresholds) all on all (N = 614 L = 31 Probability Level=0.50)	Case Estimates all on all (N = 614 L = 31 Probability Level=0.50)
Summary of item Estimates	Summary of case Estimates
Mean -0.41	Mean -0.42
SD 1.29	SD 0.89
SD (adjusted) 1.29	SD (adjusted) 0.83
Reliability of estimate 1.00	Reliability of estimate 0.87
Fit Statistics	Fit Statistics
Infit Mean Square Outfit Mean Square	Infit Mean Square Outfit Mean Square
Mean 0.91 Mean 0.92	Mean 0.94 Mean 0.93
SD 0.14 SD 0.21	SD 0.37 SD 0.58
Infit t Outfit t	Infit t Outfit t
Mean -1.30 Mean -0.74	Mean -1.17 Mean -0.09
SD 1.79 SD 1.81	SD 1.03 SD 0.77
0 items with zero scores	0 case with zero scores
0 items with perfect scores	0 case with perfect scores

Summary Results: 1993-1997 data anchored on 2000 and 2003 data

JW Items 1-40 Grades 3-10 1993+ Initial Test only (Run 6)	
Item Estimates (Thresholds) all on all (N = 902 L = 36 Probability Level=0.50)	Case Estimates all on all (N = 902 L = 36 Probability Level=0.50)
Summary of item Estimates	Summary of case Estimates
Mean 0.06	Mean 0.47
SD 1.23	SD 0.63
SD (adjusted) 1.22	SD (adjusted) 0.57
Reliability of estimate 1.00	Reliability of estimate 0.83
Fit Statistics	Fit Statistics
Infit Mean Square Outfit Mean Square	Infit Mean Square Outfit Mean Square
Mean 0.89 Mean 0.88	Mean 0.95 Mean 0.88
SD 0.22 SD 0.24	SD 0.30 SD 0.35
Infit t Outfit t	Infit t Outfit t
Mean -1.77 Mean -1.27	Mean -0.25 Mean -0.18
SD 3.71 SD 2.69	SD 1.16 SD 0.66
0 items with zero scores	0 case with zero scores
0 items with perfect scores	0 case with perfect scores

Appendix B: Sample Size, Means and Standard Deviations of all 28 samples

Grade	1993	1995	1997	2000	2002	2003
3	*n* = 322 *M* = -0.78 *SD* =0.67	*n* = 303 *M* = -0.86 *SD* = 0.61	*n* = 237 *M* = -0.51 *SD* = 0.76	*n* = 176 *M* = -0.37 *SD* = 0.75		*n* = 189 *M* = -1.25 *SD* = 0.74
5		*n* = 465 *M* = -0.30 *SD* = 0.53	*n* = 226 *M* = -0.23 *SD* = 0.50	*n* = 183 *M* = 0.37 *SD* = 0.53	*n* = 114 *M* = 0.37 *SD* = 0.47	
6	*n* = 311 *M* = -0.14 *SD* = 0.52	*n* = 337 *M* = -0.15 *SD* = 0.55	*n* = 234 *M* = -0.13 *SD* = 0.60			*n* = 174 *M* = -0.47 *SD* = 0.62
7			*n* = 314 *M* = 0.06 *SD* = 0.54	*n* = 186 *M* = 0.07 *SD* = 0.79	*n* = 102 *M* = 0.24 *SD* = 0.63	
8		*n* = 374 *M* = 0.21 *SD* = 0.58	*n* = 192 *M* = 0.13 *SD* = 0.58			
9	*n* = 392 *M* = 0.64 *SD* = 0.61	*n* = 371 *M* = 0.32 *SD* = 0.69	*n* = 105 *M* = 0.35 *SD* = 0.53	*n* = 193 *M* = 0.24 *SD* = 0.69	*n* = 135 *M* = 0.59 *SD* = 0.79	*n* = 251 *M* = 0.18 *SD* = 0.58
10			*n* = 297 *M* = 0.67 *SD* = 0.64			
11		*n* = 118 *M* = 0.96 *SD* = 0.65	*n* = 51 *M* = 0.86 *SD* = 0.80		*n* = 59 *M* = 0.71 *SD* = 0.68	

Appendix C

Comparisons of Mean Scores Across the Years for Each Grade (No adjustment for Design Effect on line 3; Effect size on line 4)

Grade 3	1995	1997	2000	2002	2003
1993	-0.78, -0.86 n=322, 303 t=-1.59, NS -0.12 (V. Small)	-0.78, -0.51 n=322, 237 t=4.48, p<.0001 0.38 (Medium)	-0.78, -0.37 n=322, 176 t=6.19, p<.0001 0.59 (Medium)		-0.78, -1.25 n=322, 189 t=-7.30, p<.0001 -0.67 (Medium)
1995		-0.86. -0.51 n=303, 237 t=5.99, p<.0001 0.51 (Medium)	-0.86, -0.37 n=303, 176 t=7.74, p<.0001 0.74 (Medium)		-0.86, -1.25 n=303, 189 t=-6.26, p<.0001 -0.59 (Medium)
1997			-0.51, -0.37 n=237, 176 t=1.76, NS 0.19 (Small)		-0.51, -1.25 n=237, 189 t=-10.02, p<.0001 -0.99 (Large)
2000					-0.37, -1.25 n=176, 189 t=-11.14, p<.0001 -1.18 (Large)
Grade 5	**1995**	**1997**	**2000**	**2002**	**2003**
1995		-0.30, -0.23 n=465, 226 t=1.72, NS 0.13 (V. Small)	-0.30, 0.37 n=465, 183 t=14.38, p<.0001 1.26 (Large)	-0.30, 0.37 n=465, 123 t=12.32, p<.0001 1.29 (Large)	
1997			-0.23, 0.37 n=226, 183 t=11.63, p<.0001 1.17 (Large)	-0.23, 0.37 n=226, 123 t=10.63, p<.0001 1.23 (Large)	
2000				0.37, 0.37 n=183, 123 t=0.02, NS 0.00	
Grade 6	**1995**	**1997**	**2000**	**2002**	**2003**
1993	-0.14, -0.15 n=311, 337 t=-0.28, NS -0.02 (V. Small)	-0.14, -0.13 n=311, 233 t=0.23, NS 0.02 (V. Small)			-0.14, -0.47 n=311, 174 t=-6.29, p<.0001 -0.59 (Medium)
1995		-0.15, -0.13 n=337, 233 t=0.47, NS 0.04 (V. Small)			-0.15, -0.47 n=337, 174 t=-5.97, p<.0001 -0.56 (Medium)
1997					-0.13, -0.47 n=233, 174 t=-5.63, p<.0001 -0.56 (Medium)
Grade 7	**1995**	**1997**	**2000**	**2002**	**2003**
1997			0.06, 0.07 n=314, 186 t=0.15, NS 0.02 (V. Small)	0.06, 0.24 n=314, 102 t=2.66, p<.004 0.23 (Small)	
2000				0.07, 0.24 n=186, 102 t=1.79, p<.04 0.32 (Small)	
Grade 8	**1995**	**1997**	**2000**	**2002**	**2003**
1995		0.21. 0.13 n=374, 192 t=-1.57, NS -0.14 (V. Small)			

Grade 9	1995	1997	2000	2002	2003
1993	0.64, 0.32 n=392, 371 t=-6.87, p<.0001 -0.49 (Medium)	0.64, 0.35 n=392, 105 t=-4.54, p<.0001 -0.49 (Medium)	0.64, 0.24 n=392, 193 t=-7.18, p<.0001 -0.63 (Medium)	-0.64, 0.59 n=392, 135 t=-0.88, NS -0.08 (V. Small)	0.64, 0.18 n=392, 251 t=-9.72, p<.0001 -0.77 (Large)
1995		0.32, 0.35 n=371, 105 t=0.39, NS 0.05 (V. Small)	0.32, 0.24 n=371, 193 t=-1.25, NS -0.12 (V. Small)	0.32, 0.59 n=371, 135 t=3.68, p<.0001 0.38 (Small)	0.32, 0.18 n=371, 251 t=-2.71, p<.007 -0.22 (Small)
1997			0.35, 0.24 n=105, 193 t=-1.36, NS -0.17 (V. Small)	0.35, 0.59 n=105, 135 t=2.66, p<.005 0.35 (Small)	0.35, 0.18 n=105, 251 t=-2.62, p<.01 -0.30 (Small)
2000				0.24, 0.59 n=193, 135 t=4.18, p<.0001 0.48 (Medium)	0.24, 0.18 n=193, 251 t=-1.11. NS -0.10 (V. Small)
2002					0.59, 0.18 n=135, 251 t=-5.84, p<.0001 -0.62 (Medium)

Grade 11	1995	1997	2000	2002	2003
1995		0.96, 0.86 n=118, 51 t=-0.90, NS -0.14 (V. Small)		0.96, 0.71 n=118, 59 t=-2.46, p<.008 -0.38 (Small)	
1997				0.86, 0.71 n=51, 59 t=-1.08, NS -0.20 (Small)	

Appendix D: Longitudinal Comparisons

Comparisons Across Years for Each Grade in the Longitudinal Samples (No adjustment for Design Effect on line 3; Effect size on line 4)

G 3, 5, 7	1995	1997	2002	2003
1993	-0.80, -0.25 $n=147$ $t=11.98$, $p<.0001$ 0.95 (Large)			
1995		-0.25, 0.14 $n=147$ $t=9.30$, $p<.0001$ 0.81 (Large)		

G 6, 8, 10	1995	1997	2002	2003
1993	-0.10, 0.25 $n=117$ $t=8.67$, $p<.0001$ 0.65 (Medium)			
1995		0.25, 0.73 $n=117$ $t=10.60$, $p<.0001$ 0.80 (Large)		

G 9, 11	1995	1997	2002	2003
1993	0.84, 0.97 $n=117$ $t=2.40$, $p<.009$ 0.20 (Small)			

G 9, 11	1995	1997	2002	2003
1995		0.67, 0.86 $n=51$ $t=2.23$, $p<.02$ 0.26 (Small)		

G 3, 9	1995	1997	2002	2003
1997				-0.28, 0.28 $n=54$ $t=5.52$, $p<.0001$ 0.79 (Large)

G 3, 5	1995	1997	2002	2003
2000			-0.35, 0.37 $n=114$ $t=12.26$, $p<.0001$ 1.11 (Large)	

G 5, 7	1995	1997	2002	2003
2000			0.34, 0.24 $n=102$ $t=-2.17$, $p<.02$ -0.18 (Small)	

G 7, 9	1995	1997	2002	2003
2000			0.13, 0.61 n=135 t=8.65, p<.0001 0.67 (Medium)	

G 9, 11	1995	1997	2002	2003
2000			0.51, 0.71 n=59 t=2.78, p<.003 0.32 (Small)	